## ABSTRACT

The two methods in present use for the analysis of minimum values, namely, a graphical method and a method of moments, are outlined and a brief discussion of each is given. In addition, a method using order statistics, devised for maximum values is adapted to be used for minimum values in the special case where the lower limit of the observed droughts is assumed to be zero.

For the general case, where the lower limit is assumed to be a positive number, a method which combines the methods of moments and order statistics is proposed. Using this method, approximate confidence bands are obtained for the predicted droughts.

# ON THE STATISTICAL ANALYSIS OF MINIMUM VALUES
# WITH APPLICATION TO DROUGHT DATA

by

Daryl Elmer Birnie

Under the direction of

Dr. Om P. Aggarwal

Department of Mathematics

University of Alberta

A THESIS

SUBMITTED TO THE SCHOOL OF GRADUATE STUDIES

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE

OF MASTER OF SCIENCE

Edmonton, Alberta. May, 1957.

ACKNOWLEDGEMENTS

I would like to express my gratitude to Dr. Om P. Aggarwal who suggested the topic and acted as supervisor during the preparation of this thesis.

I would also like to thank Dr. I. Guttman whose many suggestions were greatly appreciated.

TABLE OF CONTENTS

# CHAPTER I

# INTRODUCTION

## I. History of Statistical Theory of Extreme Values

The history of the theory of extreme values is not out of the ordinary. Different authors using different methods, independently made the same discoveries about the same time. It was a case of a common need and a similar basic knowledge to achieve the same results. Contributions were made by scientists of Russian, German, French, English and American origin.

The first work in extreme values was done in astronomy. Astronomers had to decide what to do with an outlying observation that differed greatly from the rest. Another field - gunnery - seems to be directly connected with the theory of extreme values, but there has been little or no contribution from here.

In 1922, L. von Bortkiewicz published a fundamental paper (11) on the distribution of the range and on the mean range in samples from the normal distribution as a function of the sample size. Possibly his greatest contribution was that he said that the largest normal values are new variates having distributions of their own. This, in fact was the first clear statement of the problem and also led to a new line of attack.

In 1923, R. von Mises (14) gave the first step toward a knowledge of the asymptotic distribution for normal observations, by introducing the fundamental notion of "expected largest value" (to be defined later in this chapter) which turns out to be a parameter of the asymptotic distribution.

Also in 1923, E. L. Dodd (15) studied the extreme values for distributions other than the normal and was the first to calculate the median of the extreme values. He gave formulae for Galton's distribution and the Charlier series, as well as a generalization of the normal.

The next contribution was the "Tippett's Tables", which are the numerical values of the probabilities for the extreme values from a normal distribution for different sample sizes up to one thousand, and the mean range for all the extreme values from a normal distribution from two to one thousand. These were due to L. H. C. Tippett in 1925 (16).

M. Fréchet in 1927, (17) was the first to introduce the concept of a class of initial distributions and also was the first to obtain an asymptotic distribution of the extreme values.

However, Fisher and Tippett published, in 1928 (8) the paper that is now referred to in all works on extreme values. They obtained Frechet's asymptotic distribution and constructed two other asymptotic distributions.

It should be stated here that the first researches pertaining to the theory of extreme values started with the normal distribution. This actually hampered progress due to the fact that none of the fundamental properties of extremes are related in a simple way to the normal distribution.

## 2. Aim

The aim of a statistical theory of extreme values is to explain the observed largest or smallest values arising in samples of a given size n , valid for a given period of time, or length, area, or volume, and to predict extreme values that may be expected to occur within a sample size, time, area etc.

Naturally, this prediction does not state that a definite value will occur at a particular time, but rather, it is the value that is most likely to occur within a certain interval of time, and gives limits within which the value may be expected to lie with a certain probability.

There are three essential conditions that should be fulfilled in applying statistical methods to analyse extreme value data:

(1) The variables to be considered are statistical variables.

(2) The initial distribution from which the samples are drawn, and its parameters must remain constant from one sample to the next.

(3) The observed values should be extreme values of samples of independent data.

Gumbel (reference (1)) points out that the third condition is not too critical for the following reasons:

(a) Since the actual samples used in practical applications are usually quite large, it is possible to delete a large number of the observations which may be considered to be dependent, thus leaving a sample which is still of sufficient size and that would now contain independent data.

For example, in dealing with droughts, which are defined as the minimum of the 365 daily discharges in a year, it would be possible to obtain 100 or 200 observations that would be independent.

(b) The second reason is that, as in so many other situations where the underlying causes can only be imperfectly known or assumed, the analysis of data does not wait upon the development of the most elaborate theory possible, but proceeds upon the theory built up from simple assumptions. Very often the only procedures available are

those based on independence and hence if the samples are large, it is often considered safe to proceed as though the data were actually independent.

## 3. Exact and Asymptotic Distributions for Smallest Values[1]

### (a) Exact Distributions

Let $F(x)$ be the probability that a value of the variate $X$ is less than or equal to $x$, that is, $P(X \leq x)$, and let $f(x) = F'(x)$ be the density of probability at $x$. This $f(x)$ will be referred to as the "initial distribution". Then

$$P(X \geq x) = 1 - F(x) .$$

The probability that $n$ independent observations on $X$ are all greater than $x$ is

$$\left[1 - F(x)\right]^n$$

which also gives the probability that the smallest among the $n$ independent observations is greater than $x$.

Therefore the probability that the smallest among $n$ independent observations on $X$ is less than or equal to $x$ is

$$\Phi_n(x) = 1 - \left[1 - F(x)\right]^n \tag{1.1}$$

[1] Since the main interest here is the analysis of drought data, only smallest values will be considered.

and its derivative

$$(1.2) \qquad \phi_n(x) = \Phi'(x) = n \left[ 1 - F(x) \right]^{n-1} f(x)$$

is the distribution of the smallest among $n$ independent observations. Equation (1.2) forms the basis for the whole exact theory of smallest values.

(b) Asymptotic Distributions for Smallest Values

Obviously equations (1.1) and (1.2) depend on knowledge of the initial distribution $f(x)$, which is usually not known. In order to deal with smallest values their asymptotic distributions were obtained.

An important step in the development of the asymptotic distributions was made by R. von Mises (13), who introduced the following distinction:

"A continuous variate may be either limited or unlimited in the direction of interest. If it is unlimited the moments may or may not exist. Thus there are three categories.

First those distributions which are unlimited and where all moments exist. Second, unlimited distributions where only a finite number of moments exist, and third, limited distributions."

These three categories give rise to three different types of initial distributions from which extreme values may be taken:

<u>Type I</u>: If the probability function $F(x)$ converges with increasing $x$ toward unity at least as quickly as an exponential function, then $F(x)$ is said to be of the <u>exponential type</u>. An exact definition of this type is obtained from R. von Mises' method for developing the asymptotic distribution for this type. He derived this asymptotic distribution under the condition that:

$$(1.3) \qquad \lim_{x \to \infty} \left\{ \frac{d}{dx} \left[ \frac{1 - F(x)}{f(x)} \right] \right\} = 0$$

All initial distributions possessing this property are said to be of the exponential type.

The prototype is the exponential distribution itself, while other distributions of this type are the normal and the chi-square distributions.

<u>Type: II</u>: A distribution belongs to this type if the following property is satisfied:

$$(1.4) \qquad \lim_{x \to \infty} \left[ 1 - F(x) \right] x^k = A \; ; \quad A > 0 \; ; \; k > 0$$

where $A$ is a constant, and the distribution function $F(x)$ possesses no moments of order greater than $k$. The prototype here is the Cauchy distribution , and consequently it is called the <u>Cauchy Type</u>.

Type III: If the variable x of the distribution function $F(x)$ is limited in the direction of interest, then the function $F(x)$ is said to belong to the third type.

The asymptotic distributions for these three types were found by R. A. Fisher and L. H. C. Tippett (8) in 1928. The results of this paper are given here and will be used in the main body of this thesis.

(a) For the exponential type, the asymptotic distribution of the smallest value turns out to be:

$$\Phi(x) = 1 - \exp\left[ - e^{\alpha(x-u)} \right] = 1 - \exp\left[ -e^{-y} \right] \tag{1.5}$$

$$\text{where} \quad -y = \alpha(x-u) \tag{1.5a}$$

is known as the reduced variate. $\Phi(x)$ is the probability that a drought will be more severe (that is numerically smaller) than x. The parameter u is the mode of the distribution and $\frac{1}{\alpha}$ is a scale parameter which is $\frac{\sqrt{6}}{\pi}$ times the standard deviation of the distribution.

(b) The smallest values from a distribution of the Cauchy type (Type II) have the following asymptotic distribution:

$$\pi(x) = 1 - \exp\left[ -\left(\frac{u}{x}\right)^k \right]. \qquad u < 0\ ;\ k > 0\ ;\ x \leq 0. \tag{1.6}$$

(u is not the mode here)

where the initial distribution possesses no moments of order greater than k.

(c) The third type has the variate $X$ limited by some lower limit and leads to the following asymptotic distribution:

$$(1.7) \qquad P(x) = 1 - \exp\left[-\left(\frac{x-\varepsilon}{u-\varepsilon}\right)^k\right] \qquad x \geq \varepsilon \ ; \ \varepsilon \geq 0 \ ; \ u > \varepsilon \ .$$

where $\varepsilon$ is the lower limit; $k$ is the order of the lowest derivative of the probability function that does not vanish at $x = 0$.

$P(x)$ gives the probability that an $X$ value will be less than or equal to $x$ .

5. Return Period

A concept commonly used in the treatment of smallest values is that of "return period".

If $F(x) = P(X \leq x)$ , then its reciprocal

$$(1.8) \qquad T(x) = \frac{1}{F(x)}$$

is known as the return period of $x$ . This gives the average number of observations necessary to obtain one value less than or equal to $x$ , if the observations are made at constant intervals of time.

# CHAPTER II

## 2.1 Introduction

The theory of extreme values was treated by E. J. Gumbel in a series of lectures published by the United States Bureau of Standards in February, 1954. This publication deals mainly with the analysis of largest values, for example floods, gust loads in aeronautics, etc., but states that the same method can be used for analysing smallest values.

Later in May of the same year, The American Society of Civil Engineers published a paper, also by Gumbel, which deals directly with droughts, under different basic assumptions. In this chapter, a brief outline of these two methods will be given, with examples of each.

## 2.2 Gumbel's First Method

Gumbel's first approach to the problem of analysing smallest values assumes that the initial distribution is unlimited to the left, and that the asymptotic distribution of the exponential type (see equation (1.5)) given by:

$$(2.1) \qquad F(x) = 1 - \exp\left[-e^{\alpha(x-u)}\right] = 1 - \exp\left[-e^{-y}\right] = \Phi(y)$$

where $-y = +\alpha(x-u)$ , is assumed to apply. $F(x)$ is the probability that a value of the variate $X$ will be less than or equal to $x$ . Speaking in terms of droughts, $F(x)$ gives

the probability that a future drought will be more severe (that is numerically smaller) than $x$ . $u$ and $\alpha$ are the parameters discussed in chapter I, which must be estimated.

For this distribution the return period, defined by (1.8) is given by

$$(2.2) \qquad T(x) = \frac{1}{F(x)} = \frac{1}{1 - \exp\left[-e^{y}\right]}$$

This gives the average number of observations necessary in order to obtain a drought as small as or smaller than $x$ , if the observations are made at constant intervals of time.

The method is essentially a graphical one which uses probability paper (first proposed by Powell (9)) especially designed for the treatment of extreme values. A discussion of the construction and use of probability paper in general is given in Appendix A.

In order to use this special graph paper, the observations are first ordered in decreasing magnitude and then placed on the vertical axis of the paper which is scaled linearly. The problem then arises as to the frequency at which the mth value $x_m$ should be plotted. Since the observations are ordered in decreasing magnitude, $x_m$ should be plotted at some estimate of

$$(2.3) \qquad P(X \geq x_m) = 1 - \Phi(y_m) = 1 - F(x_m)$$

Gumbel suggests that the average <u>proportion of the population $f(x)$ exceeding $x_m$</u> should be used. That is, he puts

$$(2.4) \qquad 1 - F(x_m) = 1 - \Phi(y_m) = \frac{m}{N+1}$$

where $\frac{m}{N+1}$ is the <u>expected value</u> of the proportion of the population $f(x)$ exceeding $x_m$. (This is derived in Appendix A.)

If the points $\left( \frac{m}{N+1}, x_m \right)$ are plotted on the probability paper, they should be scattered about the straight line

$$(2.5) \qquad x = u - \frac{y}{\alpha}$$

Corresponding to each observation $x_m$, there will be a return period $T(x_m)$ given by (2.2). An axis for these return periods, scaled accordingly, is included along the top of the graph paper so that the return period of any sized drought can be read directly.

The second problem arising from the use of this probability paper is that of fitting the straight line (2.5) to the plotted points. Since the relationship between $x$ and $y$ is linear, the classical method of least squares can be used.

The two regression lines $y$ on $x$ and $x$ on $y$ can both be fitted and each will give estimates of the two parameters $u$ and $\frac{1}{\alpha}$. Gumbel combines the two estimates of $u$ by taking their geometric mean. This gives

$$(2.6) \qquad \hat{u} = \bar{x} + \frac{\bar{y}_{(N)}}{\alpha}$$

which is used as the estimate of $u$. Similarly the geometric mean of the two estimates of $\frac{1}{\alpha}$ gives

$$(2.7) \qquad \frac{1}{\hat{\alpha}} = \frac{s_{(x)}}{\sigma_{(N)}}$$

as the estimate to be used for $\frac{1}{\alpha}$, where $s_{(x)}$ and $\bar{x}$ are the standard deviation and mean respectively of the sample, and $\bar{y}_{(N)}$ and $\sigma_{(N)}$ are the "theoretical" mean and standard deviation of $y$ given by[1]

$$(2.8) \qquad \bar{y}_{(N)} = \frac{1}{N} \Sigma y \quad \text{and} \quad \sigma^2_{(N)} = \overline{y^2_{(N)}} - \bar{y}^2_{(N)}$$

which are dependent on the sample size only, and have been tabulated in table II of reference (2). Using these estimates for the parameters the straight line (2.5) can be drawn on the graph.

---

[1] $\bar{y}_{(N)}$ and $\sigma_{(N)}$ are neither statistics (since they do not depend on the observations) nor purely population values (since they depend on $N$). Gumbel refers to them as the "expected" reduced mean and the "expected" reduced standard deviation.

The third problem arising from the use of this special paper is that of establishing confidence bands for the theoretical straight line. For this purpose the distribution of the mth value $x_m$ is used. Under not very restrictive conditions it can be shown that any mth value in the neighborhood of the median is asymptotically normally distributed about a mean given by (2.4) and with standard deviation

$$(2.9) \qquad \sqrt{N}\ \sigma(x_m) = \frac{\sqrt{[F(x_m)]\ [1 - F(x_m)]}}{f(x_m)}$$

where $f(x_m) = F'(x_m)$ stands for the density of probability at the value $x_m$, and is defined by (2.4). However since the approximation of the exact distribution of the mth value becomes weaker and weaker as the deviation from the median gets larger, it should be noted here that (2.9) gives valid estimates for the standard error of $x_m$ only for probabilities

$$0.15 < 1 - F(x_m) < 0.85$$

To obtain numerical values for $\sigma(x_m)$, the standard deviation of the reduced variate $y$ (which has density $\phi(y) = \Phi'(y)$) is introduced:

$$(2.10) \qquad \sqrt{N}\ \sigma(y_m) = \frac{\sqrt{[\Phi(y)]\ [1 - \Phi(y)]}}{\phi(y)}$$

which can be tabulated as a function of $y$ and has no dimension.

(See table 3.4 , reference (1)). Having these values for $\sigma(y_m)$ , $\sigma(x_m)$ can be obtained from:

$$(2.11) \qquad \sigma(x_m) = \frac{\sqrt{N}\ \sigma(y_m)}{\sqrt{N}\ \alpha}$$

To obtain the confidence bands, these values of $\sigma(x_m)$ are added to and subtracted from the theoretical values $x_m$ , situated on the straight line (2.5). This gives a probability of 0.6827 that each mth value will be contained in the interval thus obtained. If a larger probability is desired, two standard deviations are added to and subtracted from the theoretical values. This raises the probability to 0.9545.

However, as stated above, the standard errors given by (2.11) are valid only in the neighborhood of the median and hence, an extension is needed for the control curves to include the very smallest values. To do this, Gumbel utilizes the asymptotic probability distributions of the smallest and second smallest values.

If the initial distribution is of the exponential type, it has been shown (reference (18)) that the distribution of the mth largest observation (from above) converges toward

$$\phi_{N-m+1}(x_m) = \alpha_m \left[ \frac{m^m}{(m-1)!} \right] \exp \left[ - my_m - me^{-ym} \right]$$

where $y_m = \alpha_m(x_m - u_m)$ stands for the reduced variate from the population consisting of mth largest values.

Therefore the asymptotic distribution of the smallest value is

$$\phi_1(x_N) = \alpha_N \left[\frac{N^N}{(N-1)!}\right] \exp\left[-N\, y_N - Ne^{-y_N}\right]$$

and for the second smallest value is

$$\phi_2(x_{N-1}) = \alpha_{N-1} \left[\frac{(N-1)^{N-1}}{(N-2)!}\right] \exp\left[(N-1)\, y_{N-1} - (N-1)\, e^{-y_{N-1}}\right]$$

In order to extend the control curves Gumbel has shown (reference (1)) that the interval obtained by adding and subtracting the value

$$\text{(2.12)} \qquad \frac{1.1407}{\alpha_N}$$

to and from the theoretical smallest value $x_N$ situated on the straight line (2.5), will contain the observed smallest value with probability equal to 0.6827.

Similarly the interval obtained by adding and subtracting

$$\text{(2.12a)} \qquad \frac{0.7541}{\alpha_{N-1}}$$

to and from the theoretical second smallest value $x_{N-1}$ will contain the second smallest observation with the same probability.

$\alpha_N$ $(\alpha_{N-1})$ should be estimated by considering a sample made up of the smallest (second smallest) values from many samples of size $N$. However in practice this is usually not available. Gumbel uses the estimate for $\frac{1}{\alpha}$ obtained from (2.7) as the estimate for both $\frac{1}{\alpha_N}$ and $\frac{1}{\alpha_{N-1}}$.

If the points obtained by utilizing (2.12) and (2.12a) are joined with the previously obtained bands, smooth curves result and there is probability equal to 0.6827 that these curves will contain the plotted points. If a probability equal to 0.9545 is desired, the values added and subtracted to and from the smallest and second smallest values are

$$(2.13) \qquad \frac{3.0669}{\alpha} \qquad \text{and} \qquad \frac{1.7820}{\alpha}$$

respectively.

For extrapolation purposes, Gumbel has applied the principle of confidence bands to return periods. He has shown that, with probability 0.6827, a drought $x$ will occur for the first time between

$$(2.14) \qquad 0.32\ T \quad \text{and} \quad 3.13\ T$$

where $T$ is the return period corresponding to the drought $x$.

## 2.3 Example Using Gumbel's First Method

In order to illustrate the method outlined above, the following example is worked out. The drought values are observations on a certain river, call it "River R", during a 17 year period. They represent the minimum flow of water past a particular point on the river, call it "Point P" during each of the 17 years. The values, in the order in which they were observed, are given in the second column of table 2.1 .

Calculations:

(1) The observations are ordered from above and the 17 plotting positions are obtained by calculating the fractions $\frac{m}{18}$ where

$$m = 1, 2, \ldots\ldots, 17 .$$

(2) The points $\frac{m}{N+1}$ , $X_m$ are then plotted on extremal probability paper with the observed values as ordinates, and the fractions $\frac{m}{18}$ as abscissae. (See figure 2.1).

(3) In order to fit the theoretical straight line

$$x = u - \frac{y}{\alpha}$$

to the plotted points, the mean $\bar{x}$ and standard deviation $s_x$ must be calculated. Having obtained these, the estimates for the parameters $u$ and $\frac{1}{\alpha}$ are obtained from

Table 2.1 : Drought observations at Point P on River R over a 17 year period during the first quarter of each year

| Yr. | Observations (as observed) | Observations (ordered) | $1 - \Phi(y) = \frac{m}{N+1}$ | (table 1) y( ref. 3) |
|---|---|---|---|---|
| 1 | 367 | $x_1$ = 925 | 0.0556 | -1.69 |
| 2 | 358 | $x_2$ = 750 | 0.111 | -0.81 |
| 3 | 252 | $x_3$ = 605 | 0.1667 | -0.58 |
| 4 | 150 | $x_4$ = 573 | 0.222 | -0.42 |
| 5 | 605 | $x_5$ = 563 | 0.278 | -0.28 |
| 6 | 293 | $x_6$ = 430 | 0.333 | -0.10 |
| 7 | 339 | $x_7$ = 367 | 0.389 | 0.08 |
| 8 | 573 | $x_8$ = 358 | 0.445 | 0.22 |
| 9 | 750 | $x_9$ = 339 | 0.500 | 0.39 |
| 10 | 925 | $x_{10}$ = 293 | 0.556 | 0.52 |
| 11 | 563 | $x_{11}$ = 270 | 0.612 | 0.70 |
| 12 | 270 | $x_{12}$ = 252 | 0.667 | 0.90 |
| 13 | 134 | $x_{13}$ = 187 | 0.723 | 1.12 |
| 14 | 430 | $x_{14}$ = 170 | 0.777 | 1.38 |
| 15 | 170 | $x_{15}$ = 150 | 0.834 | 1.70 |
| 16 | 119 | $x_{16}$ = 134 | 0.889 | 2.15 |
| 17 | 137 | $x_{17}$ = 119 | 0.944 | 2.87 |

$$\frac{1}{\hat{\alpha}} = \frac{s_x}{\sigma_N} \quad ; \qquad \hat{u} = \bar{x} - \frac{\overline{y}_{(N)}}{\alpha}$$

where $\sigma_N$ and $\bar{y}_{(N)}$ can be obtained from table II of reference (2) .

For this example,

$$\sigma_{17} = 1.0474 \quad \text{and} \quad \overline{y(17)} = 0.5172$$

$$s_x = 224.79 \qquad \bar{x} = 381.47$$

Therefore

$$\frac{1}{\hat{\alpha}} = \frac{224.79}{1.0474} = 214.62$$

$$\hat{u} = 381.47 + (0.5172)(214.62)$$

$$= 492.47$$

and the theoretical straight line becomes

$$x = 492.47 - 215\ y$$

which is then plotted on the paper.

(4) The confidence bands are obtained by first finding $\sigma(x_m)$ for different $y_m$ values from

$$\sigma(x_m) = \frac{\sqrt{N}\ \sigma(y_m)}{\sqrt{N}\ \alpha}$$

where $\frac{1}{\alpha\sqrt{N}}$ must be calculated and $\sqrt{N}\,\sigma(y_m)$ is obtained

from table 3.4 of reference (1) , which is given here in the first three columns of table 2.2 . These values are added to and subtracted from the $x$ values situated on the straight line, that correspond to the selected $y_m$ values, for $m$ not greater than 15.

For $m = 16$ and $m = 17$ , the values

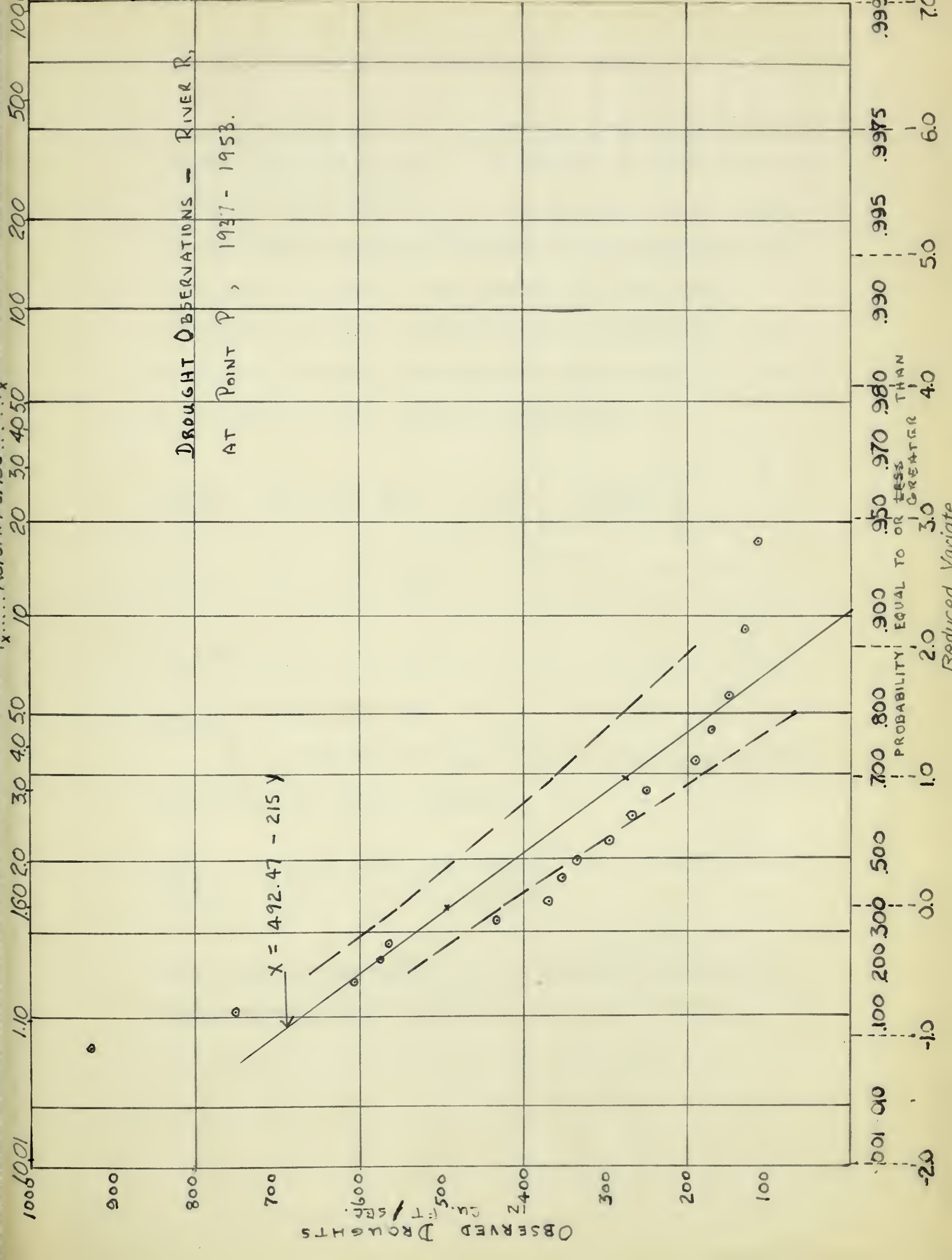
$$\frac{0.7541}{\alpha} \quad \text{and} \quad \frac{1.1407}{\alpha}$$

are added to and subtracted from the theoretical $x$ values corresponding to $y_{16}$ and $y_{17}$ . Using the estimate of $\frac{1}{\alpha}$ already obtained, these values are calculated and for this example are included in table 2.2 .

Table 2.2 Standard errors $\sigma(x_m)$ of the mth values $x_m$ , to be used as confidence band half-widths for values of $m$ up to 15. The values to be used for $m = 16$ and $m = 17$ are included in column 4 .

| y | $\sigma(y_m)\sqrt{N}$ | $\sigma(x_m) = \dfrac{\sqrt{N}\,\sigma(y_m)}{\sqrt{N}\,\alpha}$ | Confidence band half-widths for the smallest and second smallest values. |
|---|---|---|---|
| - 0.5 | 1.2431 | 59.2 | |
| 0.0 | 1.3108 | 62.2 | |
| 0.5 | 1.5057 | 71.5 | |
| 1.0 | 1.8126 | 86.2 | |
| 1.5 | 2.2408 | 106.5 | |
| 2.0 | 2.8129 | 133.7 | |
| 2.15 | | | $\frac{0.7594}{\alpha} = 163.1$ |
| 2.87 | | | $\frac{1.1407}{\alpha} = 245.5$ |

FIGURE 2.1


DROUGHT OBSERVATIONS – RIVER R,
AT POINT P, 1937 - 1953.
X = 492.47 - 215 y
OBSERVED DROUGHTS
cu. ft / sec.
PROBABILITY EQUAL TO OR GREATER THAN
Reduced Variate

## 2.4 The Second Method Proposed by E. J. Gumbel

This method takes into account the fundamental difference between floods and droughts. For droughts the lower limit must be assumed to be either zero or some positive number. In the previous method drought was treated as being unlimited to the left, which of course is unrealistic. Since the initial distribution now under consideration is a limited one in the direction of interest, the asymptotic distribution of the third type, given by (1.7) (with $k$ replaced by $\alpha$)

$$(2.15) \qquad P(X \leq x) = P(x) = 1 - \exp\left[-\left(\frac{x-\varepsilon}{u-\varepsilon}\right)^{\alpha}\right]$$

$$x \geq \varepsilon \quad ; \quad u \geq \varepsilon \quad ; \quad \alpha > 0 \quad ; \quad \varepsilon \geq 0 \, .$$

is used.

Case I : Lower limit zero

If $X$ represents drought observations and if $\varepsilon$ is taken to be zero, then (1.7) becomes:

$$(2.16) \qquad P_1(x) = 1 - \exp\left[-\left(\frac{x}{u}\right)^{\alpha}\right]$$

which gives the probability that an observed drought will be less than or equal to a particular $x$ .

The drought $x = u$ is that value that will be exceeded 36.788% of the time and Gumbel suggests that it be used to characterize a given river. It is therefore called the "Characteristic drought".

If the probability $P_1(x)$ is taken to be $\frac{1}{2}$, the median drought $\breve{x}$ can be shown to be given by

$$(2.17) \qquad \breve{x} = u(\ln 2)^{\frac{1}{\alpha}} .$$

The mode $\tilde{x}$, obtained after two differentiations of (2.16) turns out to be:

$$(2.18) \qquad \tilde{x} = u(1 - \frac{1}{\alpha})^{\frac{1}{\alpha}}$$

which is smaller than the characteristic drought $u$. Since $x$ must be positive, a mode exists only if $\frac{1}{\alpha} < 1$.

If

$$\frac{1}{\alpha} = (1 - \ln 2) = 0.30685 ,$$

then the mode $\tilde{x}$ equals the median $\breve{x}$ and the distribution is nearly symmetrical.

If $\frac{1}{\alpha} \overset{(<)}{>} 0.30685$, the mode will $\overset{\text{(exceed)}}{\text{precede}}$ the median. These facts determine the general shape of distribution (2.16) and are illustrated in figures 2.2, 2.3, 2.4.

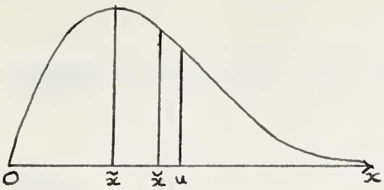



fig 2.2: $\frac{1}{\alpha} > 0.30685$

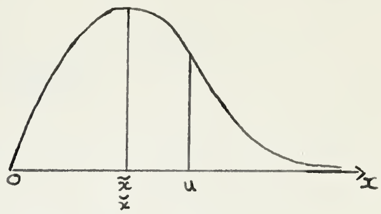



fig 2.3: $\frac{1}{\alpha} = 0.30685$

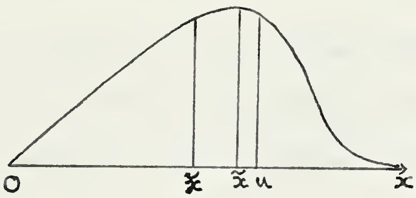



fig 2.4: $\frac{1}{\alpha} < 0.30685$

In order to analyse drought data if the lower limit is assumed to be zero, the following transformation is made in (2.16)

(2.19) Put $x = e^{z}$ and $u = e^{v}$

(2.16) now becomes:[1]

$$P_2(x) = 1 - \exp\left[-e^{\alpha(z - v)}\right]$$

(2.20) which may be written as

$$P_2(x) = 1 - \exp\left[-e^{-y_1}\right]$$

---

1 A discussion of the distribution (2.16) under a similar transformation to (2.19) is given in Chapter 3.

where

$$(2.20a) \qquad -y_1 = \alpha(z - v) = \alpha'(\log x - \log u)$$

where

$$(2.21) \qquad \alpha' = \log_e 10 \; \alpha = 2.3026 \; \alpha$$

Since the variable $y_1$ is a linear function of $\log x$ and has the same cumulative distribution function as the $y$ in (1.5), the graphical method described in the first part of this chapter may be used here as well. The only difference is that here, the common logarithms of the droughts instead of the droughts themselves are plotted against the $y_1$ values. The probabilities $P_2(x)$ and the return periods $T(x)$ given by

$$(2.22) \qquad T(x) = \frac{1}{1 - \exp\left[-\left(\frac{x}{u}\right)^{\alpha}\right]}$$

are also plotted on the extremal probability paper as before.

Instead of estimating the parameters $u$ and $\frac{1}{\alpha}$ as before, the parameters $\log u$ and $\frac{1}{\alpha'}$ are estimated. However, the estimates are obtained by the same methods and are found to be

$$(2.23) \qquad \frac{1}{\alpha'} = \frac{s(\log x)}{\sigma_{(N)}} \quad ; \quad \log u = \overline{\log x} + \frac{\bar{y}_{(N)}}{\alpha'}$$

where $\bar{y}_{(N)}$ and $\sigma_{(N)}$ are the reduced mean and reduced standard deviation obtained from (2.8) (tabulated on page 439 - 6 of reference 2).

Finally, the theoretical droughts are obtained from the graph of the straight line

$$\text{(2.24)} \qquad \log x = \log u - \frac{y}{\alpha'}$$

<u>Case 2</u>: The lower limit not equal to zero

Consider the general case, that is where the lower limit is some positive number $\varepsilon$ . Once again the cumulative distribution function

$$P(x) = 1 - \exp\left[ -\left(\frac{x - \varepsilon}{u - \varepsilon}\right)^{\alpha}\right]$$

given by (1.7) is used, where $\varepsilon$ becomes the third parameter to be estimated.

For a graphical representation of this case the transformation

$$\text{(2.25)} \qquad \log (x - \varepsilon) = \log (u - \varepsilon) - \frac{y}{\alpha'}$$

is used. However, the relationship between $y$ and $\log x$ is no longer linear. Letting $\ln x$ represent the natural logarithm of $x$ , we see that

$$\frac{d^2(\ln x)}{d\,y^2} = \frac{d}{dy}\left[\frac{1}{x}\quad\frac{dx}{dy}\right] .$$

But from (2.25)

$$x = (u - \varepsilon)\ e^{\frac{-y}{\alpha}} + \varepsilon$$

giving

$$\frac{dx}{dy} = -\frac{1}{\alpha}(u - \varepsilon)\ e^{-\frac{y}{\alpha}} = -\frac{x - \varepsilon}{\alpha}$$

Thus,

$$(2.26)\qquad \frac{d^2(\ln x)}{dy^2} = \frac{d}{dy}\left[-\frac{1}{\alpha} + \frac{\varepsilon}{\alpha\ x}\right] = \frac{\varepsilon(x - \varepsilon)}{\alpha^2\ x^2} > 0 .$$

Therefore, if $\log x$ is plotted against $y$ the resulting curve is bent downward.

Since the previous graphical estimate of the parameters is not possible, the classical method of moments is used. Differentiating (1.7), the density function $p(x)$ is obtained

$$(2.27)\qquad p(x) = \frac{\alpha}{u-\varepsilon}\left(\frac{x - \varepsilon}{u - \varepsilon}\right)^{\alpha-1} \exp\left[-\left(\frac{x - \varepsilon}{u - \varepsilon}\right)^{\alpha}\right]$$

Therefore, the kth moment of $\left(\frac{x - \varepsilon}{u - \varepsilon}\right)$ is given by

$$E\left[\left(\frac{x-\varepsilon}{u-\varepsilon}\right)^k\right] = \int_0^\infty \left[\left(\frac{x-\varepsilon}{u-\varepsilon}\right)^\alpha\right]^{k/\alpha} \exp\left[-\left(\frac{x-\varepsilon}{u-\varepsilon}\right)^\alpha\right] d\left[\left(\frac{x-\varepsilon}{u-\varepsilon}\right)^\alpha\right]$$

(2.28)

$$= \Gamma\left(1+\frac{k}{\alpha}\right)$$

Therefore, the first three moments are given by

$$\overline{\left(\frac{x-\varepsilon}{u-\varepsilon}\right)} = \Gamma\left(1+\frac{1}{\alpha}\right) ; \quad \overline{\left(\frac{x-\varepsilon}{u-\varepsilon}\right)^2} = \Gamma\left(1+\frac{2}{\alpha}\right)$$

(2.29) and

$$\overline{\left(\frac{x-\varepsilon}{u-\varepsilon}\right)^3} = \Gamma\left(1+\frac{3}{\alpha}\right) .$$

The variance $\sigma^2$ of $(x-\varepsilon)$ is

(2.30) $$\sigma^2 = (u-\varepsilon)^2 \left[\Gamma\left(1+\frac{2}{\alpha}\right) - \Gamma^2\left(1+\frac{1}{\alpha}\right)\right]$$

and the third central moment $\mu_3$ of $(x-\varepsilon)$ is

(2.31) $$\mu_3 = \overline{(x-\varepsilon)^3} - 3\overline{(x-\varepsilon)^2}\,\overline{(x-\varepsilon)} + 2\overline{(x-\varepsilon)}^3$$

Using (2.30) and (2.31), the skewness $\sqrt{\beta_1}$ can be obtained since

(2.32) $$\sqrt{\beta_1} = \mu_3 \sigma^{-3}$$

Therefore,

$$(2.32a) \quad \sqrt{\beta_1} = \frac{\Gamma(1+\frac{3}{\alpha}) - 3\Gamma(1+\frac{2}{\alpha})\Gamma(1+\frac{1}{\alpha}) + 2\Gamma^3(1+\frac{1}{\alpha})}{\left[\Gamma(1+\frac{2}{\alpha}) - \Gamma^2(1+\frac{1}{\alpha})\right]^{+\frac{3}{2}}}$$

This expression depends only on $\frac{1}{\alpha}$ and hence if $\sqrt{\beta_1}$ is replaced by the sample value $\sqrt{b_1}$ , an estimate of $\frac{1}{\alpha}$ can be obtained. ($\sqrt{\beta_1}$ are tabulated in reference 2 for different values of $\frac{1}{\alpha}$).

To estimate $u$ , (2.30) is used. The relationship

$$(2.33) \quad u = \bar{x} + \sigma A(\alpha)$$

where

$$(2.33a) \quad A(\alpha) = \left[1 - \Gamma(1+\frac{1}{\alpha})\right]\left[\Gamma(1+\frac{2}{\alpha}) - \Gamma^2(1+\frac{1}{\alpha})\right]^{-\frac{1}{2}}$$

is obtained, and since $\frac{1}{\alpha}$ has already been estimated an estimate of $u$ can be obtained if

$$s = (\overline{x^2} - \bar{x}^2)^{\frac{1}{2}}$$

is used as an estimate of $\sigma$ .

To estimate $\varepsilon$ , equation (2.30) is written in the form

$$\varepsilon = \frac{\bar{x} - u\Gamma(1+\frac{1}{\alpha})}{1 - \Gamma(1+\frac{1}{\alpha})}$$

and $\bar{x}$ is replaced by its value from (2.33) giving

$$(2.34) \qquad \varepsilon = u - \sigma B(\alpha)$$

where

$$(2.34a) \qquad B(\alpha) = \left[ \Gamma(1 + \frac{2}{\alpha}) - \Gamma^2(1 + \frac{1}{\alpha}) \right]^{-\frac{1}{2}}$$

and is also tabulated in reference 2 . The estimates of $u$ and $\frac{1}{\alpha}$ already obtained are used along with the sample standard deviation $s$ for $\sigma$ .

A criterion as to whether the lower limit should be taken as zero or not is established from equations (2.33) and (2.34) since

$$\varepsilon \geq 0 \quad \text{if} \quad \bar{x} + s \left[ A(\alpha) - B(\alpha) \right] \geq 0 \ .$$

A more convenient form of this condition is

$$(2.35) \qquad \varepsilon \geq 0 \quad \text{if} \quad \frac{\overline{x^2}}{\bar{x}^2} \leq \frac{\Gamma(1 + \frac{2}{\alpha})}{\Gamma^2(1 + \frac{1}{\alpha})}$$

If the equality is fulfilled within the limits of random sampling the lower limit is taken to be zero. If $\varepsilon$ turns out to be negative but small, it can safely be assumed to be zero.

After the three parameters have been estimated, the theoretical droughts are obtained from (2.25) and may be plotted against $y$ on logarithmic extremal probability paper[1]. The expected droughts for the desired return periods can easily be read off the graph.

## 2.5 Example of Drought Analysis Using the Method of Moments

Case 1:

The lower limit $\varepsilon$ is assumed to be zero. Table 2.3 gives the droughts observed at Point P on River R over a 17 year period, their logarithms and the frequencies at which they are to be plotted.

If logarithmic probability paper is available the droughts themselves are plotted against the frequencies $\frac{m}{N+1}$, as in figure 2.5. If ordinary extremal probability paper is being used, the logarithms of the droughts are plotted, and the theoretical straight line

$$\log x = \log u - \frac{y}{\alpha'},$$

is fitted to the plotted points.

The estimates of $\log u$ and $\frac{1}{\alpha'}$ are obtained as solutions of (2.23)

---

1 This paper differs from ordinary extremal probability paper in that the axis corresponding to the drought values is scaled logarithmically.

$$\frac{1}{\alpha'} = \frac{s(\log x)}{\sigma_N} \quad ; \quad \log u = \overline{\log x} + \frac{\overline{y}_{(N)}}{\alpha'}$$

For this example:

$$\overline{\log x} = 1.9667 \; ; \; s(\log x) = \overline{(\log x)^2} - (\overline{\log x})^2 = 2.005$$

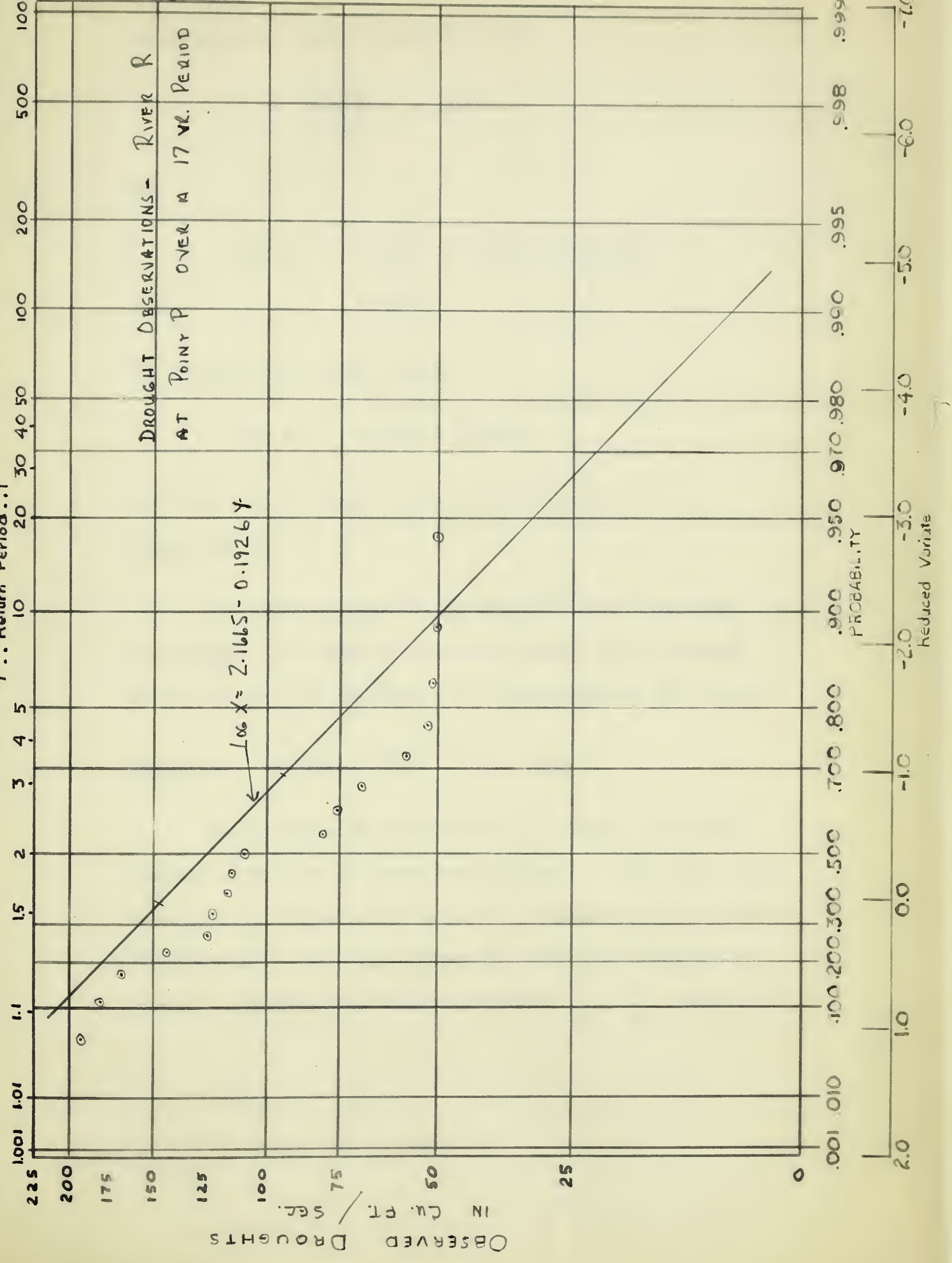
$\sigma_N$ and $\overline{y}_N$ are obtained from table II of reference (2) to be

$$\sigma_{17} = 1.0411 \quad ; \quad \bar{y}_{17} = 0.5181$$

Table 2.3: Droughts observed at Point P on River R over a 17 year period

| Yr. | Droughts x (as obsv.) | x (ordered) | Log. x (ordered) | $\frac{m}{N + 1}$ m = 1,..17 |
|---|---|---|---|---|
| 1 | 76 | 189 | 2.2765 | 0.0556 |
| 2 | 57 | 182 | 2.2601 | 0.111 |
| 3 | 51 | 169 | 2.2279 | 0.167 |
| 4 | 50 | 142 | 2.1523 | 0.222 |
| 5 | 182 | 123 | 2.0899 | 0.278 |
| 6 | 189 | 122 | 2.0864 | 0.333 |
| 7 | 123 | 115 | 2.0607 | 0.389 |
| 8 | 108 | 113 | 2.0531 | 0.444 |
| 9 | 142 | 108 | 2.0334 | 0.500 |
| 10 | 169 | 80 | 1.9031 | 0.556 |
| 11 | 113 | 76 | 1.8808 | 0.611 |
| 12 | 68 | 68 | 1.8325 | 0.667 |
| 13 | 115 | 57 | 1.7559 | 0.722 |
| 14 | 122 | 52 | 1.7160 | 0.777 |
| 15 | 52 | 51 | 1.7076 | 0.833 |
| 16 | 50 | 50 | 1.6990 | 0.889 |
| 17 | 80 | 50 | 1.6990 | 0.944 |
| | | 1747 | 33.4342 | |

Fig: 2.5


Drought Observations - River R
At Point P over a 17 yr. Period
Log X = 2.1665 - 0.1926 Y
T . . Return Period . T
1.001 1.01 1.1 1.5 2 3 4 5 10 20 30 40 50 100 200 500 1000
Probability
.001 .010 .100 .200 .300 .500 .700 .800 .900 .950 .970 .980 .900 .995 .998 .999
Reduced Variate
2.0 1.0 0.0 -1.0 -2.0 -3.0 -4.0 -5.0 -6.0 -7.0
225 200 175 150 125 100 75 50 25 0
Observed Droughts in Cu. Ft. / Sec.

Therefore, the required estimates are

$$\frac{1}{\alpha'} = \frac{0.2005}{1.0411} = 0.1926$$

and

$$\log u = 1.9667 + (0.5181)(0.1926)$$
$$= 2.1665$$

The straight line then becomes

$$\log x = 2.1665 - 0.1926\, y$$

and this line is drawn on the probability paper as in figure 2.5 .

The return period of any drought can be read from this graph. For example the return period of the drought 25 cu. ft./sec. is obtained to be approximately 29.5 years.

Case 2: The lower limit $\varepsilon$ not zero.

The droughts measured in River R during the second quarter of each of 17 years are analysed. Table 2.4 gives the observed values and some of the preliminary calculations required, and Table 2.5 gives the remaining calculations needed to estimate the three parameters $u$, $\frac{1}{\alpha}$, and $\varepsilon$ .

Table 2.4: Droughts observed at Point P on River R during the second quarter of each year over a 17 year period.

| Yr. | x | $x^2$ | $x^3$ |
|---|---|---|---|
| 1 | 126 | 15876 | 2,000,376 |
| 2 | 164 | 26896 | 4,410,944 |
| 3 | 115 | 13225 | 1,520,875 |
| 4 | 139 | 19321 | 2,685,619 |
| 5 | 375 | 140,625 | 52,734,375 |
| 6 | 238 | 56,644 | 13,481,272 |
| 7 | 176 | 30,976 | 5,451,776 |
| 8 | 238 | 56,644 | 13,481,272 |
| 9 | 343 | 117,649 | 40,353,607 |
| 10 | 339 | 114,921 | 38,958,219 |
| 11 | 218 | 47,524 | 10,360,232 |
| 12 | 113 | 12,769 | 1,442,897 |
| 13 | 174 | 30,276 | 5,268,024 |
| 14 | 282 | 79,524 | 22,425,768 |
| 15 | 103 | 10,609 | 1,092,727 |
| 16 | 149 | 22,201 | 3,307,949 |
| 17 | 118 | 13,924 | 1,643,032 |
| | 3410 | 809,604 | 220,618,964 |

Table 2.5: Estimate of the Three Parameters: River R, Point P

( 1.) Mean drought $\bar{x} = \frac{3410}{17} = 200.59$

( 2.) Mean square $\overline{x^2} = \frac{809,604}{17} = 47,623.76$

( 3.) Variance $S^2 = \overline{x^2} - \bar{x}^2 = (47,623.76)-(200.59)^2 = 7,387.42$

( 4.) St. Dev. $S = 85.95$

( 5.) $S^3 = 634,948.75$

( 6.) $\bar{x}^3 = 220,618,964 \div 17 = 12,977,585$

( 7.) $m_3 = \bar{x}^3 - 3(\bar{x}^2)(\bar{x}) + 2\bar{x}^3 = 461,009.8$

( 8.) Skewness $\sqrt{b_1} = m_3 s^{-3} = 0.7261$

( 9.) $\frac{1}{\hat{\alpha}}$ : From table IV reference (2) = 0.5357

(10.) $\frac{1}{\hat{\alpha}'} = (0.4343)(0.5357) - 0.2327$

(11.) $A(\alpha)$ : table IV reference (2) = 0.2269

(12.) $\hat{u} = \bar{x} + sA(\alpha) =$

$= 200.59 + (85.95)(0.2269) = 220.09$

(13.) $B(\alpha)$ : table IV reference (2) = 2.0248

(14.) $sB(\alpha) = \hat{u} - \hat{\varepsilon} = (85.95)(2.0248) = 174.03$

(15.) $\hat{\varepsilon} = \hat{u} - sB(\alpha) = 2.20.09 - 174.03 = 46.06$

# CHAPTER III

## Analysis of Minimum Values Using Order Statistics.

### 3.1 Introduction

In January 1954, the National Advisory Committee for Aeronautics in the United States published a paper by Julius Lieblein (reference (4)) outlining an entirely different method for analysing extreme value data. However the method is given only for maximum values since their main concern was gust loads on an airplane in flight. In this chapter this method, which is one of order statistics, will be outlined for use in the analysis of minimum values; in particular, droughts where the lower limit is assumed to be zero.

### 3.2 The method of order statistics.

Let $X$ represent drought values. Then the probability of a drought more severe than $x$ (that is, numerically smaller than $x$) is given by (1.7) with $\varepsilon$, the lower limit, assumed to be zero

$$(3.1) \qquad P(X \leq x) = G(x) = 1 - \exp\left[-\left(\frac{x}{u}\right)^{\alpha}\right]$$

$$0 \leq x < \infty \; ; \quad u > 0 \; ; \quad \alpha > 0 \, .$$

However, the method outlined by Lieblein is based on the assumption that the observed data are independent observations from a statistical distribution of the form of (1.5)

$$P(X \leq x) = F(x) = \exp\left[-e^{-\alpha(x-u)}\right] = \exp\left[-e^{-y}\right]$$

where $y = \alpha\,(x-u)$; $\alpha > 0$; $-\infty < x < \infty$ .

If the following transformation is made in (3.1),

$$(3.2) \qquad \left(\frac{X}{u}\right)^{\alpha} = e^{-Z} \qquad \text{or} \qquad Z = \alpha(-\ln X + \ln u)$$

then

$$P(X \leq x_1) = P(u\,e^{-\frac{Z}{\alpha}} \leq x_1) = P(\ln u - \frac{Z}{\alpha} \leq \ln x_1)$$

$$(3.3) \qquad = P\left[-Z \leq \alpha(\ln x_1 - \ln u)\right] = P\left[Z \geq \alpha(-\ln x_1 + \ln u)\right]$$

$$= P\,(Z \geq z_1) \,.$$

That is, the probability of a drought $X$ being less than or equal to $x_1$ , is equivalent to the probability of a $Z$ value being greater than or equal to the corresponding $z_1$ .

The cumulative distribution function of the new variate $Z$ is given by

$$(3.4) \qquad P(Z \leq z) = \exp\left[-e^{-z}\right]$$

where

$$(3.4a) \qquad z = \alpha(-\ln x + \ln u)$$

which is precisely the form of the distribution function (1.5) considered by Lieblein with $y$ replaced by $z$, $x$ by $-\ln x$, and $u$ by $-\ln u$.

Therefore, if the negative logarithms of the droughts are considered instead of the droughts themselves the method outlined by Lieblein can be applied directly.

First, a combination of the two parameters to be estimated

$$\xi = -\ln u + \frac{z}{\alpha} \tag{3.5}$$

is introduced. Although the distribution is completely specified by the two parameters $-\ln u$ and $\frac{1}{\alpha}$, it will be shown that the quantity $\xi$ makes it possible to estimate them simultaneously and not as two separate parameters.

If the probability $P(Z \leq z)$ is chosen to be some fixed value, then the corresponding $z$ value can be obtained from relationship (3.4) (tabulated in reference (3)). Having obtained this $z$ value, the corresponding value of $\xi$ can be obtained from (3.5). That is, $P$ having been fixed, the values of $z$ and $\xi$ are automatically fixed. To denote this dependence of $z$ and $\xi$ on $P$, they will be written

$$z_p \quad \text{and} \quad \xi_p \quad .$$

If $P$ is chosen at different levels, say $P = 0.10$, $0.05$, $0.01$, etc., the corresponding $\xi_p$'s will be the estimates used for the predictions for the negative logarithms of the droughts, such that smaller droughts will occur only 10, 5, 1, etc. times respectively in 100 future droughts.

It is by the proper choice of $P(Z \leq z)$ that estimates of the parameters $-\ln u$ and $\frac{1}{\alpha}$ are obtained from the value of $\xi_p$. If $P$ is chosen to be $\frac{1}{e} = 0.36788$, it is evident from (3.4) that $z_p = 0$. Putting $z_p = 0$ in equation (3.5) $\xi_p$ is seen to be

$$(3.6) \qquad \xi_p = -\ln u + \frac{z_p}{\alpha} = -\ln u$$

which gives the required estimate for $-\ln u$. Similarly, if the limiting value of $P$ is considered, that is if we let $P$ approach one, the corresponding values of $\xi_p$ and $z_p$ become indefinitely large, but their ratio

$$(3.7) \qquad \xi_p^* = \frac{\xi_p}{z_p} = \frac{-\ln u}{z_p} + \frac{1}{\alpha}$$

may be considered to be a new parameter which approaches $\frac{1}{\alpha}$.

From the above discussion, it is evident that the solutions of both the problems of estimation and prediction are embodied in the one quantity

$$\xi_p = -\ln u + \frac{z_p}{\alpha}$$

and estimation of this quantity will be the main problem dealt with here. The method of attack will be that of order statistics.

If the values in a sample of $N$ observations are arranged in say, increasing order of magnitude, that is,

$$x_1 \leq x_2 \leq \cdots\cdots \leq x_N ,$$

then these $x_i$'s are called <u>order statistics</u>.

Here, the observations are the negative logarithms of drought values and they must first be ordered in increasing magnitude, such that

$$-\ln x_1 \leq -\ln x_2 \leq \cdots\cdots \leq -\ln x_N .$$

The aim is to determine the weights $w_i$, $i = 1,..,n$, for all the $n$ order statistics so that the linear estimator

$$L = \sum_{i=1}^{N} w_i(-\ln x_i) \tag{3.8}$$

has the following properties:

(i) The mathematical expectation of $L$ equals the parameter to be estimated.

That is,

$$E(L) = \xi_p \tag{3.9}$$

This condition makes $L$ an unbiased estimator.

(ii) The mean square error (MSE), which in this case is the same as the variance, is as small as possible, consistent with condition (i).

That is

$$(3.10) \qquad \text{MSE}\ (L) = \sigma^2\ (L) = E\left[L - E(L)\right]^2 = \text{a minimum.}$$

For each value $-\ln x_i$, there corresponds a $z_i$ of the following form (from (3.4a))

$$z_i = \alpha(-\ln x_i + \ln u)$$

Therefore,

$$(3.11) \qquad E(-\ln x_i) = -\ln u + \frac{1}{\alpha} E(z_i)$$

and consequently

$$(3.12) \qquad E(L) = \sum_{i=1}^{N} w_i \left[E(-\ln x_i)\right] = \sum_{i=1}^{N} w_i \left[-\ln u + \frac{1}{\alpha} E(z_i)\right]$$

$$= \xi_p = -\ln u + \frac{1}{\alpha} z_p$$

This is required to be an identity and hence if the coefficients of $-\ln u$ and $\frac{1}{\alpha}$ are equated, the conditions on the weights $w_i$, are obtained as follows:

$$\sum_{i=1}^{N} w_i = 1$$

(3.13) and

$$\sum_{i=1}^{N} E(z_i)\ w_i = z_p$$

The values $E(z_i)$ have been tabulated in reference (5) .

Turning to the variance, there is obtained

$$(3.14) \qquad Var(L) = \sum_{i=1}^{N} w_i^2 \sigma_{-\ln x_i}^{2} + \sum_{j=1}^{N} \sum_{\substack{i=1 \\ i \neq j}}^{N} w_i w_j \sigma_{(-\ln x_i)(-\ln x_j)}$$

From the definition of $-\ln x_i$ in terms of $z_i$ and utilizing the properties of variances and covariances of linear estimators and then making a simplification in notation:

$$\sigma_{-\ln x_i}^{2} = \left(\frac{1}{\alpha}\right)^2 \sigma_{z_i}^{2} = \left(\frac{1}{\alpha}\right)^2 \sigma_i^{2}$$

(3.15) and

$$\sigma_{(-\ln x_i)(-\ln x_j)} = \left(\frac{1}{\alpha}\right)^2 \sigma_{z_i z_j} = \left(\frac{1}{\alpha}\right)^2 \sigma_{ij}$$

whence,

$$(3.16) \qquad V_N = Var.(L) = \left[ \sum_{i=1}^{N} \sigma_i^{2} w_i^2 + \sum_{\substack{i=1 \\ i \neq j}}^{N} \sum_{j=1}^{N} \sigma_{ij} \; w_i w_j \right] \left(\frac{1}{\alpha}\right)^2$$

$$= \text{a minimum subject to } (3.9) .$$

This is a constrained minimum problem for variation in the unknown $w_i$ and is equivalent to finding the unconstrained minimum of:

$$(3.17) \qquad G_1 = \left( \sum_i \sigma_i^{\,2} w_i^2 + \sum_i \sum_{\substack{j \\ i \neq j}} \sigma_{ij} \; w_i w_j \right) \left( \frac{1}{\alpha} \right)^2$$

$$+ \lambda_1 \left( \sum_i w_i - 1 \right) + \mu_1 \left( \sum_i E(z_i) w_i - z_p \right)$$

where $\lambda_1$ and $\mu_1$ are Lagrange multipliers. This is the same as minimizing

$$(3.18) \qquad G = \alpha^2 G_1 = \sum_i \sigma_i^{\,2} w_i^2 + \sum_i \sum_{\substack{j \\ i \neq j}} \sigma_{ij} \; w_i w_j$$

$$+ \lambda \left( \sum_i w_i - 1 \right) + \mu \left( \sum_i E(z_i) w_i - z_p \right)$$

since $\frac{1}{\alpha}^2 > 0$ is a constant, though unknown. Setting the derivative with respect to $w_k$ equal to zero

$$(3.19) \qquad 2 \sigma_k^{\,2} w_k + \sum_{\substack{i=1 \\ i \neq k}}^{N} \sigma_{ik} \; w_i + \lambda + \mu \, E(z_k) = 0 \qquad k = 1, 2 \ldots, N$$

(3.19) is a system of $N$ linear equations which if combined with (3.13) form a simultaneous system of $N + 2$ equations in the $N + 2$ unknowns

$$w_1, w_2, \ldots\ldots, w_N, \lambda \text{ and } \mu .$$

Before the sets of equations (3.13) and (3.19) can be solved, the coefficients $E(z_k)$, $\sigma_k^{\,2}$ and $\sigma_{ik}$ must be

determined. As previously stated the values of $E(z_k)$ are tabulated in reference (5). The variances and covariances $\sigma_k^2$ and $\sigma_{ik}$ involve complicated integrals which Lieblein has expressed in terms of simpler ones which are tabulated in reference (6). These mean, variance, and covariance values are combined into one table -- table III of reference (4) -- for values of $n$ up to and including $n = 6$.

This table gives the coefficients in the equations (3.13) and (3.19). The right hand sides of these $n + 2$ equations are

$$1, z_p, 0, 0, \ldots\ldots, 0,$$

and the solutions

$$w_i, \quad \lambda \text{ and } \mu \quad ,$$

are linear combinations of these with numerical coefficients which involve only $\sigma_i^2$, $\sigma_{ij}$ , and $E(z_i)$ but not $z_p$ .

Therefore the solutions are all of the form:

$$w_i = a_i + b_i z_p$$

$$\lambda = c_1 + d_1 z_p \qquad i = 1, 2, \ldots\ldots, N . \tag{3.20}$$

$$\mu = c_2 + d_2 z_p$$

Substituting these values of $w_i$ in equations (3.13) and (3.19) yields a solution for the minimum variance of the following form:

$$(3.21) \qquad V_{N,min} = (A_N z_p^2 + B_N z_p + C_N)\left(\frac{1}{a}\right)^2$$

The quantities $a_i$ and $b_i$ for the weights, and the coefficients $A_N$, $B_N$, $C_N$, of $V_{N,min}$ are all given in table one, reference (4) for $N = 2$ to $N = 6$. The procedure for samples larger than 6 is explained in reference (4) and is outlined in Appendix B of this thesis. Having obtained the weights $w_i$, the $\xi_p$ estimates for different probabilities P can be calculated as illustrated in the example given in section 3.3.

Lieblein has made an important extension in his work with extreme values by including methods by which information concerning the mean and variance of the estimator $\xi_p$ can be obtained. The mean value of an estimator indicates whether on the average the estimate given is too high or too low relative to the parameter estimated. The variance makes it possible to compare the performances of different estimators by indicating how much the estimators scatter among themselves; that is, it is a basis for constructing a measure of efficiency of the estimator.

In order to have a standard of comparison, all variances are scaled by dividing them into a theoretically specified variance $Q_{LB}$ which is known as the "Cramér - Rao Lower Bound"

-- (reference (7), pg. 480). This variance is less than or equal to the variance of any unbiased estimator of the parameter in question.

The resulting efficiency is an absolute number between 0 and 1, and is given by

$$(3.22) \qquad \text{Efficiency } (L) = E_N(L) = \frac{Q_{LB}}{Q_N}$$

where $L$ is the estimator and $Q_N = V_{N,min}$. The quantities $E_n$, which depend on $z_p$ (since $Q_N$ depends on $z_p$) and consequently on $P$, are tabulated for $N = 2$ to $N = 6$, for different probability levels $P$, in table III reference (4).

Lieblein uses the standard deviations of the estimator $\xi_p$ to establish confidence limits around the predicted values. For a fixed probability $P$, the interval

$$(3.23) \qquad \hat{\xi}_p \pm \text{(one standard deviation)}$$

will contain the true unknown parameter

$$\bar{\xi}_p = -\ln u + \frac{z_p}{\alpha}$$

about 68% of the time. If two standard deviations are used, the percentage rises to 95%.

## 3.3 An example of drought analysis using order statistics.

The drought observations given in table 3.1 were obtained

at Point P on River R over a 17 year period. Since there are 17 observations, they must be split into subgroups according to the rules given in Appendix B. Three subgroups are obtained, two consisting of 6 observations each and a third consisting of 5 observations. The negative logarithms of the droughts are obtained and these are ordered within each subgroup in increasing magnitude. The remaining calculations are presented in the form of two self explanatory work sheets suggested by Lieblein.

Table 3.1 : Drought observations on River R and their negative logarithms.

| Yr. | Droughts x | -ln x | Yr. | Droughts | -ln x |
|---|---|---|---|---|---|
| 1 | 76 | -4.3307 | 10 | 169 | -5.1299 |
| 2 | 57 | -4.0431 | 11 | 113 | -4.7274 |
| 3 | 51 | -3.9318 | 12 | 68 | -4.2195 |
| 4 | 50 | -3.9120 | 13 | 115 | -4.7449 |
| 5 | 182 | -5.2040 | 14 | 122 | -4.8040 |
| 6 | 189 | -5.2418 | 15 | 52 | -3.9512 |
| 7 | 123 | -4.8122 | 16 | 50 | -3.9120 |
| 8 | 108 | -4.6821 | 17 | 80 | -4.3820 |
| 9 | 142 | -4.9958 | | | |

Work Sheet 1 :

1. Subgroup sizes and proportionality factors.

$N = 17 = km + m' = 2 \times 6 + 5$

$t = \frac{km}{N} = \frac{12}{17} = 0.7059$ $\qquad$ $t' = \frac{5}{17} = 0.2941$

$\frac{t^2}{k} = 0.2492$ $\qquad$ $(t')^2 = 0.0865$

$k = 2 \quad ; \quad m = 6 \quad ; \quad m' = 5 \ .$

2. (a) Main subgroups

Weights $a_i$ and $b_i$ (from table 1, reference (4))

| i | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $a_i$: | 0.3555 | .2255 | .1656 | .1211 | .0835 | .0489 |
| $b_i$: | -0.4593 | -.0360 | .0732 | .1267 | .1495 | .1458 |

-ln $x_i$ in increasing order

| | $-\ln x_1$ | $-\ln x_2$ | $-\ln x_3$ | $-\ln x_4$ | $-\ln x_5$ | $-\ln x_6$ |
|---|---|---|---|---|---|---|
| 1: | -5.2418 | -5.2040 | -4.3307 | -4.0431 | -3.9318 | -3.9120 |
| 2: | -5.1299 | -4.9958 | -4.8122 | -4.7274 | -4.6821 | -4.2195 |

$$\bar{T} = \frac{\sum_{i=1}^{6} a_j x_i}{k} + \frac{\sum_{i=1}^{6} b_i x_i}{k} z_p = -4.8401 + 0.4386\, z_p$$

(b) Remainder group

Weights $a_i'$ and $b_i'$ (from table 1 reference (4))

| i | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $a_i'$: | 0.4189 | 0.2463 | 0.1676 | 0.1088 | 0.0584 |
| $b_i'$ | -0.5031 | 0.0065 | 0.1305 | 0.1817 | 0.1845 |

Work Sheet 1 (Cont'd.)

$-\ln x'_i$ in increasing order

| $-\ln x'_1$ | $-\ln x'_2$ | $-\ln x'_3$ | $-\ln x'_4$ | $-\ln x'_5$ |
|---|---|---|---|---|
| -4.8040 | -4.7449 | -4.3820 | -3.9512 | -3.9120 |

$$T' = \sum_{i=1}^{5} a_i x'_i + \sum_{i=1}^{5} b'_i x'_i z_p = -4.5738 + 0.3745 z_p$$

Therefore

$$\hat{\xi}_p = t\bar{T} + t'T'$$

$$\hat{\xi}_p = (0.7059)(-4.8401 + 0.4386 z_p) + (0.2941)(-4.5738 + 0.3745 z_p)$$

$$= 4.7618 + 0.4197 z_p$$

and the estimates for $-\ln u$ and $\frac{1}{\alpha}$ are:

$$-\ln u = -4.7618 \quad \text{and} \quad \frac{1}{\alpha} = 0.4197$$

WORK SHEET 2:

| P | $z_p$ | $\hat{\xi}_p$ Predicted Values of -ln x | $Q_m = Q_6$ Table III ref. 4 | $Q_{m'} = Q_5$ Table III ref. 4 | $\mathrm{Var}(\hat{\xi}_p) = \frac{t^2}{x} Q_m + (t')^2 Q_{m'}$ | 68% conf. half-width $\sigma(\hat{\xi}_p) = \sqrt{\mathrm{Var.}(\hat{\xi}_p)}$ | $Q_{LB} = \frac{Q_0}{m}$ ($Q_0$ from Table III ref. 4 ) | Efficiency $E = \frac{Q_{LB}}{\mathrm{Var}(\hat{\xi}_p)}$ |
|---|---|---|---|---|---|---|---|---|
| 0.36788 | 0 | -4.7618 | $0.1912 \frac{1}{\alpha^2}$ | $0.2314 \frac{1}{\alpha^2}$ | $0.06766 \frac{1}{\alpha^2}$ | 0.1092 | $0.06392 \frac{1}{\alpha^2}$ | 0.945 |
| 0.50 | 0.3665 | -4.6080 | $0.2319 \frac{1}{\alpha^2}$ | $0.2787 \frac{1}{\alpha^2}$ | $0.08190 \frac{1}{\alpha^2}$ | 0.1117 | $0.08110 \frac{1}{\alpha^2}$ | 0.990 |
| 0.90 | 2.2504 | -3.8173 | $1.0077 \frac{1}{\alpha^2}$ | $1.2283 \frac{1}{\alpha^2}$ | $0.3556 \frac{1}{\alpha^2}$ | 0.2503 | $0.3144 \frac{1}{\alpha^2}$ | 0.884 |
| 0.95 | 2.9702 | -3.5152 | $1.5417 \frac{1}{\alpha^2}$ | $1.9035 \frac{1}{\alpha^2}$ | $0.5488 \frac{1}{\alpha^2}$ | 0.3109 | $0.4705 \frac{1}{\alpha^2}$ | 0.857 |
| 0.99 | 4.6002 | -2.8311 | $3.2723 \frac{1}{\alpha^2}$ | $4.0706 \frac{1}{\alpha^2}$ | $1.1676 \frac{1}{\alpha^2}$ | 0.4535 | $0.9611 \frac{1}{\alpha^2}$ | 0.823 |
| 0.999 | 6.9073 | -1.8628 | $6.9204 \frac{1}{\alpha^2}$ | $8.6517 \frac{1}{\alpha^2}$ | $2.4729 \frac{1}{\alpha^2}$ | 0.6560 | $1.9802 \frac{1}{\alpha^2}$ | 0.801 |
| 1.00 | | 0.4197 | $0.1320 z_p^2 \frac{1}{\alpha^2}$ | $0.1667 z_p^2 \frac{1}{\alpha^2}$ | $0.04731 \frac{1}{\alpha^2} z_p^2$ | | $0.03576 z_p^2 \frac{1}{\alpha^2}$ | 0.756 |

## 3.4 The general case where the lower limit is not zero.

Let X be a random variable representing drought values. The probability that a drought more severe (that is numerically smaller) than x will occur is given by

$$(3.24) \qquad P(X \leq x) = P(x) = 1 - \exp\left[ - \left(\frac{x - \varepsilon}{u - \varepsilon}\right)^{\alpha} \right]$$

$$\varepsilon \leq x \leq \infty \ ; \quad u > \varepsilon \ ; \quad \alpha > 0 \ ; \quad \varepsilon > 0 \ .$$

where $\varepsilon$, $u$, and $\frac{1}{\alpha}$ are parameters which must be estimated.

If this case is to be treated using the method of order statistics outlined in section 3.2, (3.24) must be put in the form

$$P(X \leq x) = F(x) = \exp\left[ -e^{-\alpha(x - u)} \right] = \exp\left[ -e^{-y} \right]$$

The transformation linking these two cumulative distribution functions is given by

$$\left(\frac{X - \varepsilon}{u - \varepsilon}\right)^{\alpha} = e^{-Z}$$

(3.25)

$$\text{or} \qquad Z = \alpha \left[ -\ln(x - \varepsilon) + \ln(u - \varepsilon) \right]$$

in (3.24) .

An effort was made using order statistics to obtain a method that would yield unbiased estimates simultaneously for the three parameters $u$, $\frac{1}{\alpha}$, $\varepsilon$, but this was unsuccessful. Instead, the following "combined" method is proposed to handle the case where the lower limit is not zero.

First the lower limit $\varepsilon$ is estimated by the method of moments outlined in section 2.4. If this estimate of $\varepsilon$ is then subtracted from each of the original observations on $X$, the cumulative distribution function (3.24) reduces to one of the form (3.1) in the new variable, say

$$X_{(1)} = X - \varepsilon$$

and the two parameters $u_{(1)} = u - \varepsilon$, and $\frac{1}{\alpha}$.

Table (3.3) gives the estimate of $u$ and $\frac{1}{\alpha}$ obtained by applying this combined method to drought observations on River R over a period of 17 years. Since more than one set of data was desired, the years were split up into quarters and the droughts during each quarter were analysed. For the purpose of comparison the estimates of $u$ and $\frac{1}{\alpha}$ obtained by the method of moments on the same data are also given in table (3.3).

Table 3.3

| Quarter | Method of Moments | | Combined Method | |
|---|---|---|---|---|
| | $u$ | $\frac{1}{\alpha}$ | $u$ | $\frac{1}{\alpha}$ |
| 1 | 425.7 | 0.5904 | 410.3 | 0.4026 |
| 2 | 220.1 | 0.5357 | 226.0 | 0.5611 |
| 3 | 116.2 | 0.4684 | 117.1 | 0.4665 |
| 4 | 142.8 | 0.9048 | 142.3 | 0.7516 |

One of the rather serious disadvantages in applying the method of moments to the case where $\varepsilon$ is not equal to zero is that confidence intervals for the predicted droughts are extremely difficult to obtain; in fact there is no method available at this time by which they can be obtained. This disadvantage is partially overcome if the combined method is used, as approximate confidence intervals can be obtained for the predicted values of $-\ln(x - \varepsilon)$ and these can be converted into approximate confidence limits for the actual predicted values.

Table 3.4 gives the predicted droughts with return periods 10, 20, and 100 years (denoted by $x_{10}$, $x_{20}$, and $x_{100}$ respectively) both for the method of moments and the combined method. In addition the confidence band half-widths for the predictions of $-\ln(x_n - \varepsilon)$ $(n = 10, 20, \text{ or } 100)$ obtained by the combined method are given along with the confidence limits for

the predicted droughts. It must be kept in mind, that these confidence limits are only approximate, since there is no control on the estimate used for $\varepsilon$.

Table 3.4

| n | $-\ln(x_n - \varepsilon)$ | Predicted Values: Method of Moments | Predicted Values: Combined Method | Approx. 68% conf. band half-widths for $-\ln(x-\varepsilon)$ | Approx. 68% conf. limits for the predicted $x_n$ |
|---|---|---|---|---|---|
| 1st Quarter | | | | | |
| 10 | -4.8806 | 117.6 | 142.7 | 0.2938 | 109.3-187.7 |
| 20 | -4.5260 | 81.6 | 103.6 | 0.3649 | 75.3-144.3 |
| 100 | -3.7230 | 38.5 | 52.5 | 0.5323 | 35.4- 81.6 |
| 2nd Quarter | | | | | |
| 10 | -3.9268 | 96.8 | 96.6 | 0.3336 | 82.6-117.0 |
| 20 | -3.5229 | 81.0 | 80.0 | 0.4157 | 68.5- 97.4 |
| 100 | -2.6083 | 60.9 | 59.7 | 0.6063 | 53.5- 71.0 |
| 3rd Quarter | | | | | |
| 10 | -3.6331 | 45.4 | 46.7 | 0.2782 | 37.6- 58.9 |
| 20 | -3.2973 | 35.2 | 36.0 | 0.3456 | 28.0- 47.1 |
| 100 | -2.5369 | 21.3 | 21.5 | 0.5041 | 16.5- 29.8 |
| 4th Quarter | | | | | |
| 10 | -3.0492 | 42.2 | 49.0 | 0.4483 | 41.4- 61.0 |
| 20 | -2.5082 | 35.5 | 40.2 | 0.5568 | 34.9- 49.4 |
| 100 | -1.2831 | 29.7 | 31.5 | 0.8121 | 29.5- 36.0 |

CHAPTER IV

## Conclusion

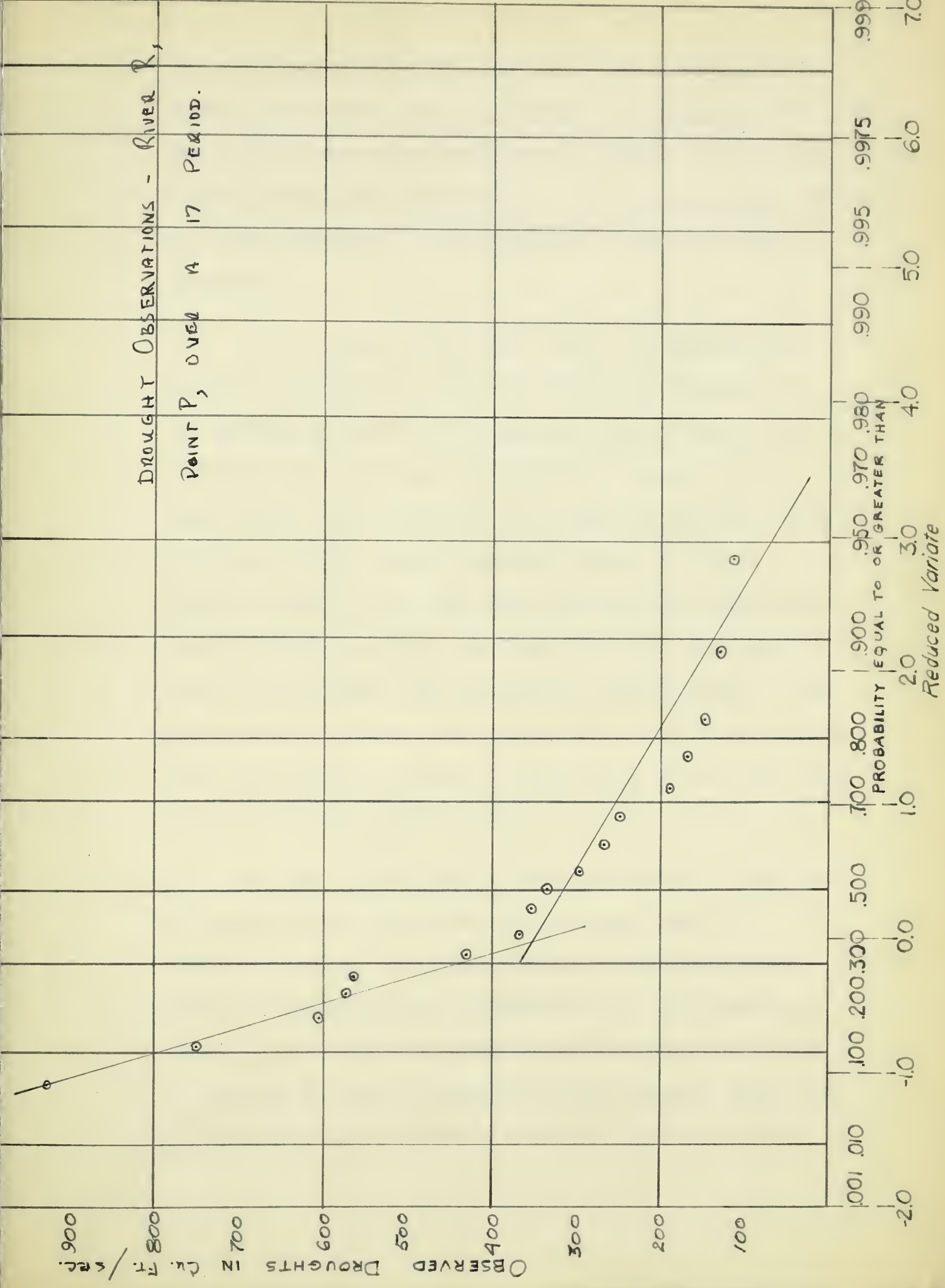
Four methods have been presented to deal with the problem of analysing minimum values. The first was a graphical method which utilized a special probability paper; the second was based on the classical method of moments; the third used order statistics to deal with the special case where the lower limit was assumed to be zero; and the fourth combined the methods of moments and order statistics to handle the general case where the lower limit is assumed to be some positive number. In this chapter a brief discussion of these four methods will be given.

The graphical method presented in section 2.2, although very simple and compact, has one rather serious disadvantage. The plotted points do not tend to cluster around one straight line as Gumbel claims they will. Rather they seem to form two lines as is illustrated in figure 4.1, which is a graph of the same observations used in figure 2.1. The uppermost line in figure 4.1 is interpreted as being formed by moderate droughts which do not belong to the extreme-values proper, but still to the initial distribution. The second line is formed by more severe droughts and only it can be used for extrapolation purposes. This discontinuity leads to a loss of about 37% [1] of the information furnished by the observations.

---

[1] This percentage is quoted by Gumbel in reference 2. It appears that all the observations larger than the "characteristic drought" u (see section 2.4) have to be discarded.

FIGURE 4.1


DROUGHT OBSERVATIONS - RIVER R,
POINT P, OVER A 17 PERIOD.
OBSERVED DROUGHTS IN CU. FT./SEC.
900
800
700
600
500
400
300
200
100
PROBABILITY EQUAL TO OR GREATER THAN
.001 .010 .100 .200 .300 .500 .700 .800 .900 .950 .970 .980 .990 .995 .9975 .999
Reduced Variate
-2.0 -1.0 0.0 1.0 2.0 3.0 4.0 5.0 6.0 7.0

The method of moments, the second method proposed by Gumbel, corrects this loss of information by recognizing the fact that one must assume droughts to be extreme values from a limited (to the left) distribution. This of course, gives rise to a third parameter -- the lower limit -- which has to be estimated.

As can be seen by equation (2.34) the estimate given by the method of moments for the lower limit is dependent on the standard deviation of the sample so that the smaller the variation within the sample the larger the estimate for the lower limit. Due to this fact it is quite possible for a river with very low (more severe) observed droughts to yield a higher estimate for the lower limit than one with higher (less severe) observed droughts. The estimated lower limit may even turn out to be larger than the smallest observed value. If the latter value is reliable, the method fails. Another possibility which would cause the method to fail would be for the lower limit to take on a large negative value.

The third method, that of order statistics, is outlined in chapter III for application when the lower limit is assumed to be zero. A comparison between this method and the method of moments is given in reference (4). The comparison is carried out after first combining the estimates for $u$ and $\alpha$ given by the method of moments into one estimator (which will be referred to as the "moment's estimator") that has similar

form to that of the order statistics estimator $\hat{\xi}_p$ given by relationship (3.5). The main interest is to compare the efficiency of the two estimators. In order to obtain these efficiencies, the first two moments of the sample mean and standard deviation and the covariance of the mean and standard deviation must be obtained. For the moment's estimator, only the first two moments of the sample mean are readily obtainable by standard procedures. Therefore, the comparison is carried out using a simplified form of the moment's estimator which is valid only for large samples. However it is shown that the original moments estimator is much less efficient than the one considered. From this comparison the following advantages of the order statistics estimator seem apparent.

(a) The method of order statistics provides an estimator known to be unbiased, whose efficiency can be simply and accurately evaluated.

(b) The estimator is more efficient than a simplified form of the moment's estimator, for samples of about 20 or more and probability $P = 0.95$ and more.

(c) The order statistics method uses a more exact procedure to obtain the reliability of predicted values, and this procedure yields smaller confidence intervals in many cases.

The following two limitations on the method of order statistics should be noted.

(a) The method is applicable only when the assumptions on which it is based are considered to be approximately satisfied; namely, the observations constitute an independent sample from the population

$$F(x) = \exp\left[-e^{-a(x-u)}\right]$$

(b) The method of order statistics treats each observation on an individual basis, and hence is not very suitable for large samples since they cannot be grouped.

The combined method outlined in section 3.4 is a rather obvious combination of the methods of moments and order statistics. However it has the advantage that, for the first time, confidence limits are obtainable for the predicted droughts, even though they are approximate. The predicted values compare very well with those obtained by the method of moments with the exception of those for the first quarter (see table 3.3). As stated previously in this chapter the method of moments estimate for the lower limit depends on the variation within the observed sample, and consequently the predicted values tend to be either too high -- for a small sample variation -- or too low for a large sample variation. Since the variation within the first sample is rather large (see table 2.1) it would seem logical (based on the above discussion) to expect the predicted droughts for this quarter, by the method of moments, to be too low. As can be seen by

table 3.3 all the predicted values for the first quarter calculated by the combined method are higher than those obtained by using the method of moments. Since the predicted values for the other three quarters, where the variations are not abnormally high, are quite comparable, it would seem that this combined method tends to give more accurate estimates for samples with large variations, although the degree of accuracy is not known and further investigation is needed on this point.

A natural extension of this investigation into the analysis of minimum values would be to obtain a method that would give unbiased estimates for $u$, $\frac{1}{\alpha}$, and $\varepsilon$ simultaneously in such a way that the efficiency of the estimators could be obtained without too much difficulty. An attempt was made to accomplish this by applying the method of maximum likelihood and also by trying to extend the method of order statistics, but both were unsuccessful. However there is certainly scope for further investigation into both these methods.

## APPENDIX A

### Probability Paper

Let $X$ be a continuous random variable, unlimited in both directions and having cumulative distribution function

$$P(X \leq x) = F(x)$$

Assume the existence of a linear transformation

$$(A.1) \qquad x = \mu + \beta y$$

where $\mu$ and $\beta$ are location and scale parameters respectively, both having the dimension of $x$. The new variable $y$, known as the reduced variable, has dimension zero. A well known example of such a transformation is used in standardizing a normal variate by putting

$$(A.2) \qquad z = \frac{x - \mu}{\sigma}$$

where $\mu$ and $\sigma$ are the population mean and standard deviation respectively.

If the values of the variable $X$ are plotted against the reduced variable $y$, a straight line would naturally result since the relationship between them is linear. However, the problem arises as to how the $y$ value corresponding to a particular $x$ is arrived at, since $\mu$ and $\beta$ are parameters that are unknown.

Corresponding to the cumulative distribution function $F(x)$ of $X$, there is a cumulative distribution function of $y$, say $\Phi(y)$ and

$$\Phi(y) = F(x) . \tag{A.3}$$

The important point here is that $\Phi(y)$ is independent of the parameters $\mu$ and $\beta$. Therefore, if an estimate of $\Phi(y)$ could be obtained, the corresponding $y$ value would automatically be known. To estimate $\Phi(y)$ an estimate of $F(x)$ is obtained from the observations on $X$ as follows:

Let $$x_1 \leq x_2 \leq \cdots\cdots\cdots \leq x_N ,$$

be $N$ observations on the variate $X$, assumed to have the cumulative distribution function $F(x)$, ordered in increasing magnitude. Then the mth value $x_m$ has the density function

$$g(x_m) = \frac{N!}{(N-m)!\,(m-1)!}\left[\int_{-\infty}^{x_m} f(x)\,dx\right]^{m-1}\left[\int_{x_m}^{\infty} f(x)\,dx\right]^{N-m} f(x_m) \tag{A.4}$$

Let $\xi_m$ be the proportion of the population $f(x)$ preceding $x_m$, that is,

$$\xi_m = \int_{-\infty}^{x_m} f(x)\,dx \tag{A.5}$$

Clearly $1 > \xi_m > 0$.

Then the density function of $\xi_m$ is

$$(A.6) \qquad \psi(\xi_m) = \frac{N!}{(N-m)!\,(m-1)!} (\xi_m)^{m-1} (1 - \xi_m)^{N-m}$$

The expected value of $\xi_m$ is

$$(A.7) \qquad E(\xi_m) = \frac{N!}{(N-m)!\,(m-1)!} \int_0^1 (\xi_m)^m (1 - \xi_m)^{N-m} \, d\,\xi_m$$

$$= \frac{m}{N+1}$$

That is, the average proportion of the population $f(x)$ preceding the mth value $x_m$, is $\frac{m}{N+1}$, and this average proportion is taken as the estimate used for $\Phi(y)$. Hence, corresponding to each observation $x_m$, there is an estimate of the cumulative distribution function $F(x) = \Phi(y)$, and therefore an estimate of $y$.

Probability paper is a rectangular grid on which the variate $X$ is plotted on one of the axes - usually the vertical - on a linear scale. The other axis is scaled in such a way that if the estimates for the cumulative distribution function $F(x)$ are plotted against the $x$'s, a straight line will result. This enables one to obtain the theoretical straight line

$$x = \mu + \beta y$$

$\mu$

and to estimate the parameters $\mu$ and $\beta$ (by ordinary regression procedures) without ever actually obtaining the $y$ values.

If the observations on X are ordered in decreasing magnitude as in section 2.2, that is

$$x_1 \geq x_2 \geq \cdots\cdots\cdots \geq x_N$$

the same probability paper can be used if the proportion estimated is $1 - F(x)$ instead of $F(x)$. As an estimate for this quantity, the average proportion of the population $f(x)$ exceeding the mth value (from above) is used. This average is found to be $\frac{m}{N+1}$, which is used in section 2.2.

As an example of the use of probability paper, consider a variate X distributed normally with mean $\mu$ and standard deviation $\sigma$. The cumulative distribution function of this variate is given by

$$\text{(A.8)} \qquad P(X \leq x_1) = F(x_1) = \int_{-\infty}^{x_1} \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] dx$$

The linear transformation here is

$$\text{(A.9)} \qquad z = \frac{x-\mu}{\sigma}$$

and (A.8) becomes

$$\text{(A.10)} \qquad \Phi(z_1) = \int_{-\infty}^{z_1} \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{z^2}{2}\right] dx$$

which is free of parameters and has been tabulated.

If one considers a sample of N observations from this distribution, ordered in decreasing magnitude

$$-\infty \leq x_N \leq x_{N-1} \leq \cdots\cdots \leq x_1 \leq \infty$$

then the value $\frac{m}{N + 1}$ , which is the expectation of the proportion of the population exceeding $x_m$ , can be used as an estimate for $1 - F(x)$ . If the points $\left( \frac{m}{N + 1} , x_m \right)$ are plotted on normal probability paper they should cluster around the straight line

$$x = \hat{\mu} + \hat{\beta} z$$

where $\hat{\mu}$ and $\hat{\beta}$ are estimates of the population mean and standard deviation, and are obtained by ordinary regression procedures.

# APPENDIX B

## Extension to larger samples.

Most samples are larger than the trivial size of six. The following will outline how these larger samples are to be handled. The principle is to treat them as sets of subgroups of six (or five). Two cases arise:

Case I : Sample size an exact multiple of 5 or 6.

Let the sample size $n = km$, where $m$ is the size of the subgroup; and $k$ is the number of subgroups in the sample. Now each subgroup is treated as a separate sample of size $m$. This is legitimate if the original sample is divided into subgroups in such a way that each subgroup consists of statistically independent observations.

From each subgroup a "subestimator" is formed:

(B.1) $$T_i = \sum_{j=i}^{m} w_j x_j \qquad i = 1, 2, \ldots\ldots\ldots, k .$$

where the weights $w_j$ are obtained as in chapter III and are the same for each subgroup of size $m$. The arithmetic mean of these $k$ subgroup estimators $T_i$ is then taken to be the grand sample estimator:

(B.2) $$\bar{T} = \frac{1}{k} \sum_{i=1}^{k} T_i$$

The variance of $\bar{T}$ is given by

$$(B.3) \qquad \text{Var. } (T) = \frac{1}{k} Q_m$$

since this variance is that of a mean of $k$ independent quantities, each of which has the same variance $Q_m$ (given in table III, reference 4).

The efficiency of $\bar{T}$ is, since $n = km$ and the $T_i$'s and therefore $\bar{T}$, are unbiased:

$$(B.4) \qquad \text{Eff.} = \frac{Q_{LB}}{\text{Var}(\bar{T})} = \frac{\frac{1}{km} Q_o}{\frac{1}{k} Q_m} = \frac{\frac{1}{m} Q_o}{Q_m} = E_m$$

where $Q_{LB}$ is the ~~Cremar~~-Rao lower bound, which can be obtained from table III, reference 4. Since the efficiency depends only on the size $m$ of the subgroup and increases with increased $m$, the largest size of subgroup should be chosen if there is a choice. Cramer

Case II : Sample size not an exact multiple of 5 or 6.

The aim of course, is to establish as simple rules as possible without too great a loss in efficiency. Actually two separate cases arise:

(a) For $n = 7$ up to large values:

(i) Use the partition $n = 6k + m'$ if the remainder $m' = 2, 3, 4,$ or $5$. If $m' = 1$, use $5k + m''$.

(ii) If $n$ is a multiple of 30 plus 1, that is $n = 31, 61, 91$, etc., write

$n = 30k + 1 = (30k - 5) + 6 = 5(6k - 1) + 6$

that is, split the sample into $6k - 1$ subgroups of 5 and a remainder subgroup of 6.

In order to obtain the estimator $\hat{\xi}_p$ and its variance, assume the sample has been split into two parts, one consisting of $k$ equal subgroups of size $m$, and the other consisting of the remainder subgroup of size $m'$. The average $\bar{T}$ of the first $k$ subgroups is found as outlined in case I. Then a subestimator $T'$ is found from the remainder subgroup by using the weights $w_i'$ for a sample of size $m'$, that is

$$(B.5) \qquad T' = \sum_{i=1}^{m'} w_i' \, x_i' \,.$$

Finally, a weighted average of $\bar{T}$ and $T'$ is found and this is taken as the final estimator $\hat{\xi}_p$.

$$(B.6) \qquad \hat{\xi}_p = t\bar{T} + t'T'$$

where

$$(B.6a) \qquad t = \frac{km}{n} \quad \text{and} \quad t' = \frac{m'}{n} = 1 - t$$

Since all the subgroups are independent, and hence $\bar{T}$ and $T'$, and since the variance of the mean is $\frac{1}{k} Q_m$, therefore,

$$\text{(B.7)} \qquad \text{Var.}\ (\hat{\xi}_p) = \frac{t^2}{k} Q_m + (t')^2 Q_{m'}$$

The efficiency can be obtained in the same way as outlined in case I.

(b) <u>n extremely large</u>:

If the number of subgroups is of the order 50 to 1000, the amount of computation becomes very laborious. The following short cut method is suggested to deal with these cases. Although there is quite a large loss in efficiency, the method is of practical value in as much as a loss in efficiency is effectively a loss in sample size, which is not too important if an extensive amount of data is available.

First arrange all $n$ observations in order of increasing size, and then rank them from one to $n$. Select the three observations $x_r$ whose ranks are the nearest integers to $0.03n$, $0.20n$ and $0.85n$. These will be denoted by:

$$x_{0.03n}, \quad x_{0.20n}, \quad \text{and} \quad x_{0.85n}.$$

The predicted values $\hat{\xi}_p$, for various probability levels $P$, can be computed from

(B.8) $$\hat{\xi}_p = x_{0.20n} + 0.3256\ (Z_p + 0.4759)(x_{0.85n} - x_{0.03n})$$

(See ref. 4)

The variance of this estimator can be computed from:

(B.9) $$\sigma^2(\hat{\xi}_p) = 8.6916\ d^2 - 0.0681\ d + 1.5442$$

where $d = 0.3256\ Z_p + 0.1549$ .

BIBLIOGRAPHY


(1) E. J. Gumbel, Statistical theory of extreme values and some practical applications, National Bureau of Standards Applied Mathematics Series 33 (Feb. 12, 1954).

(2) E. J. Gumbel, Statistical theory of droughts, American Society of Civil Engineers Proceedings, Vol. 80, Separate No. 439 (May, 1954).

(3) National Bureau of Standards, Probability tables for the analysis of extreme-value data, Applied Math. Series 22, (July 6, 1953).

(4) Julius Lieblein, A new method of analysing exteme-value data, National Advisory Committee for Aeronautics, Technical note 3053, January 1954.

(5) National Applied Mathematics Labor~~i~~atories, Table of the first moment of ranked extremes project 550 - 39 NACA and National Bureau of Standards (Sept. 20, 1951).

(6) Julius Lieblein, On the exact evaluation of the variances and covariances of order statistics in samples from the extreme-value distribution, Ann. Math. Stat., Vol. 24, No. 2, June 1953, pp. 282 - 287.

(7) Harold Cramér, Mathematical methods of statistics, Princeton University Press (1946).

(8) R. A. Fisher and L. H. C. Tippett, Limiting forms of the frequency distribution of the largest or smallest member of a sample, Proc. Cambridge Phil. Soc. 24, Pt. 2, 180 - 190 (April, 1928).

(9) R. W. Powell, A simple method of estimating flood frequencies, Civ. Eng. 105 - 6 (1943).

(10) L. von Bortkiewicz, Variationsbreite und kleinen Zahlen, p. 23 (Leipzig, B. G. Teubner, 1898).

(11) L. von Bortkiewicz, Variationsbreite und mittlerer Fehler, Sitzungsberichte d. Berliner Math. Gesellschaft 21, 3 - 11 (1922).

(12) J. Fourier, Règle usuelle pour la recherche des résultats moyens d'un grand nombre d'observations, Bul. Sci. Math. Astron. Phys. et Chim II, 88 - 90 (Paris, 1824).

(13) R. Helmert, Ueber den Maximalfehler einer Beobachtung, Zeitschrift f. Vermessungswesen 6, 131 - 147 (1877).

(14) R. von Mises, La distribution de la plus grande de n valeurs, Revue Math. L'Union Interbalcanique 1, 1 - 20 (Athens, 1936).

(15) F. L. Dodd, The greatest and the least variate under general laws of error, Trans. Am. Math. Soc. 25, 525 - 539 (1923).

(16) L. H. C. Tippett, On the extreme individuals and the range of samples taken from a normal population, Biometrika 17, Pts. 3 and 4, 364 - 387 (1925).

(17) M. Fréchet, Sur la loi de probabilité de l'écart maximum, Ann. Soc. Polonaise Math. Cracovie 6, 93 - 116 (1927).

(18) E. J. Gumbel, Les valeurs extrêmes des distributions statisques, Ann. de l'Institut Henri Poincaré V, 115 - 158 (1935).